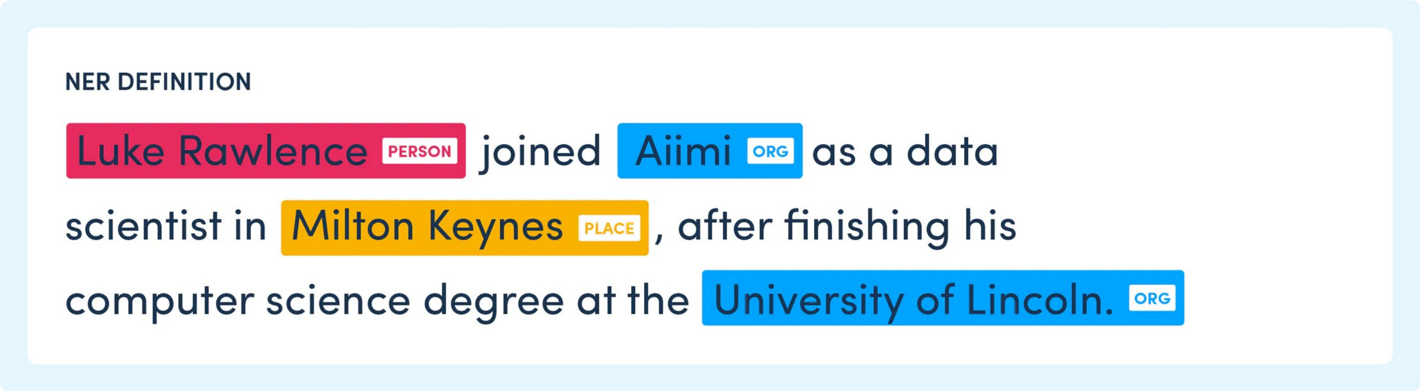
Contextualized Language Model-based Named Entity Recognition in Slovak Texts

supervisor: Mgr. Endre Hamerlik
consultant: Mgr. Marek Šuppa
student: Bc. Dávid Šuba

Named Entity Recognition (NER)

- recognition and categorization of named entities in text
- Person, Location, Organization, Product, Time ……
- one of the base tasks in NLP (applications: information extraction from CV, recommender systems…)



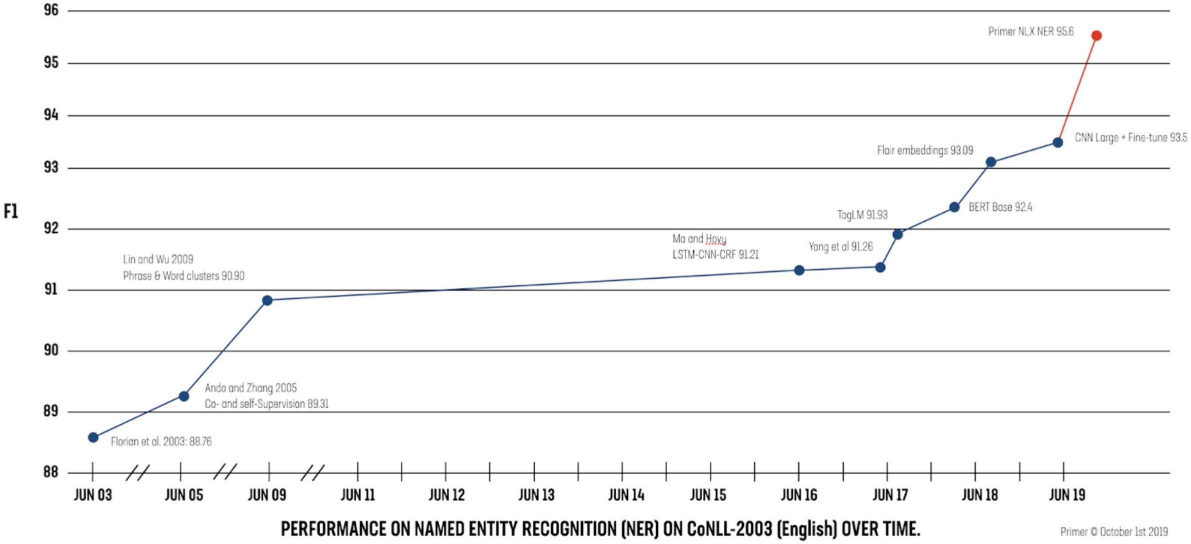
NER DEFINITION

Luke Rawlence PERSON joined Aiimi ORG as a data scientist in Milton Keynes PLACE, after finishing his computer science degree at the University of Lincoln. ORG

source: https://www.aiimi.com/insights/aiimi-labs-on-named-entity-recognition

Motivation

- state-of-the-art are contextualized language models based on deep neural networks
- the best systems for NER in slovak texts are rule based and are far behind [1]
- the main reason is lack of resources and datasets



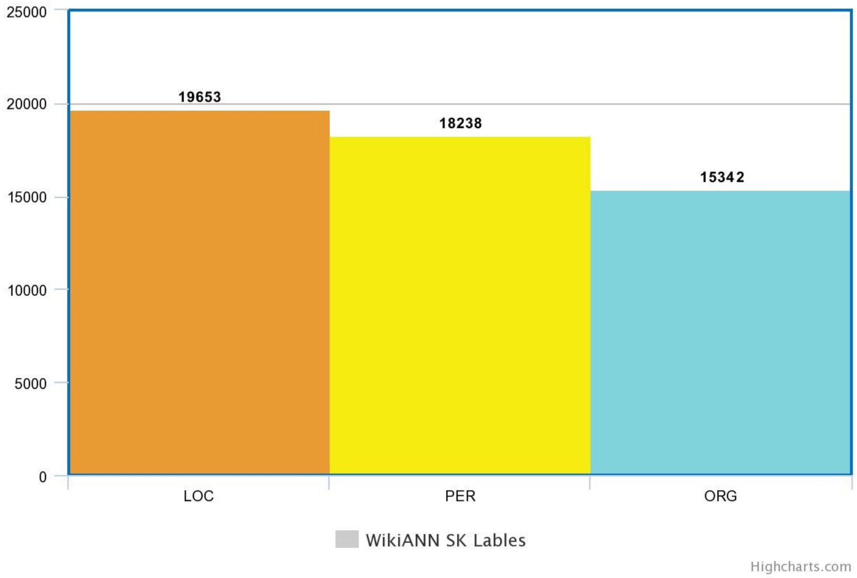
source: https://company.primer.ai/blog/a-new-state-of-the-art-for-named-entity-recognition/
ww.aiimi.com/insights/aiimi-labs-on-named-entity-recognition

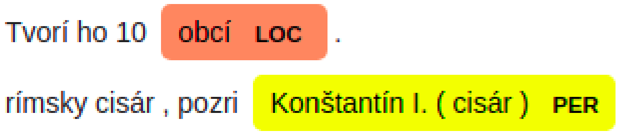
Goals

- utilize, improve and expand existing slovak NER datasets
- use large multilingual language models and transfer learning for named entity recognition in slovak texts

WikiANN (PAN-X)

- automatically generated ‘silver standard’ NER dataset for 282 languages from Wikipedia articles [3]
- utilizes markups and knowledge-base links included in articles in cross-lingual tagging
- LOC, PER, ORG tags

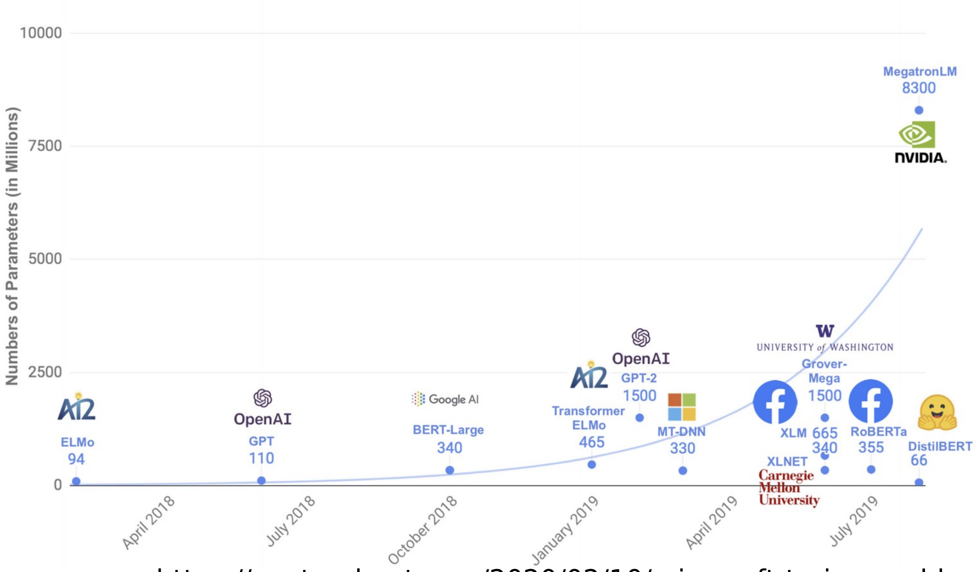



Tvorí ho 10 obcí LOC .

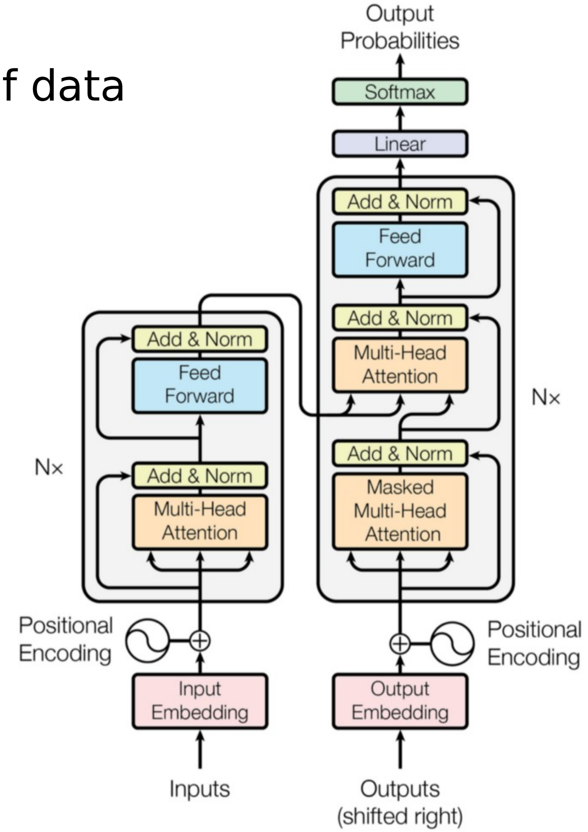
rímsky cisár , pozri Konštantín I. (cisár) PER

Transformers language models

- BERT (Google), GPT-2/3 (OpenAI), RoBERTa (Facebook), ALBERT (Google)
- deep-learning models with attention layers
- state-of-the-art standard
- trained unsupervised on huge amounts of data
- fine-tunning for specific tasks (transfer learning)



source: https://venturebeat.com/2020/02/10/microsoft-trains-worlds-largest-transformer-language-model/



source: [2]

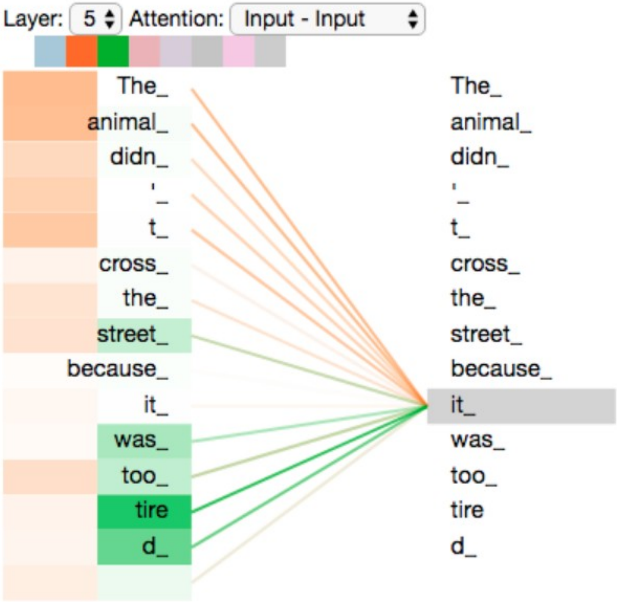
Transformers language models

- attention: the attention-mechanism allows model to decide at each step of sequence which other parts of the sequence are important
- Q – query vector (word in a sequence)
- K – keys (all words in sequence)
- V – values (all words in sequence)

$$\mathrm{Attention}(Q,K,V) = \mathrm{softmax}(\frac{QK^T}{\sqrt{d_k}})V$$

- positional encoding: allows model to utilize order of sequence

$$PE_{(pos,2i)} = sin(pos/10000^{2i/d_{\mathrm{model}}})$$
$$PE_{(pos,2i+1)} = cos(pos/10000^{2i/d_{\mathrm{model}}})$$



Layer: 5 Attention: Input - Input

source: http://jalammar.github.io/illustrated-transformer/

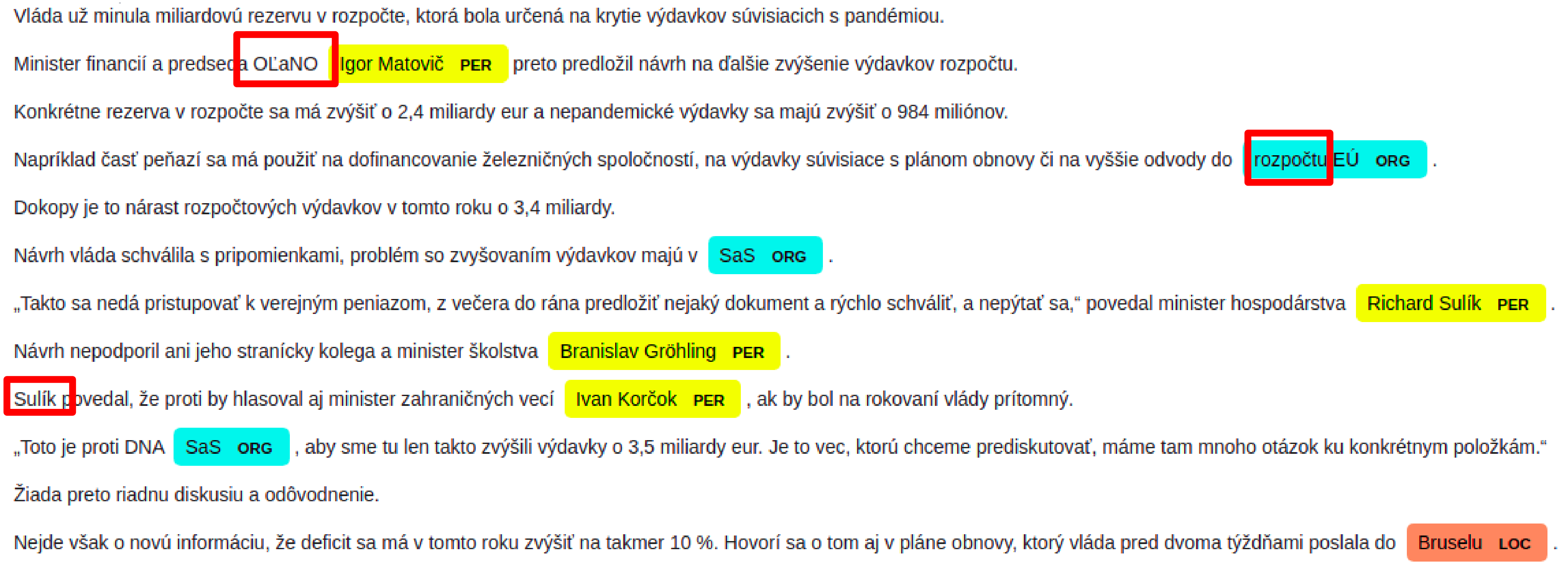
Trankit + WikiANN

- trankit – python tool for NLP tasks based on transformers
- most pipelines relies at XML-Roberta – big multilanguage model trained at 2.5TB of data from CommonCrawl dataset (https://commoncrawl.org/)
- for every language and component (NLP task) is excluded specific adapter (weights across layers), which is after-trained at specific dataset

Precision	Recall	F1
88.9	88.51	88.7

- results at ‘gold‘ dataset would be markantly lower

Trankit + WikiANN



Vláda už minula miliardovú rezervu v rozpočte, ktorá bola určená na krytie výdavkov súvisiacich s pandémiou.

Minister financií a predseda OĽaNO Igor Matovič PER preto predložil návrh na ďalšie zvýšenie výdavkov rozpočtu.

Konkrétne rezerva v rozpočte sa má zvýšiť o 2,4 miliardy eur a nepandemické výdavky sa majú zvýšiť o 984 miliónov.

Napríklad časť peňazí sa má použiť na dofinancovanie železničných spoločností, na výdavky súvisiace s plánom obnovy či na vyššie odvody do rozpočtu EÚ ORG .

Dokopy je to nárast rozpočtových výdavkov v tomto roku o 3,4 miliardy.

Návrh vláda schválila s pripomienkami, problém so zvyšovaním výdavkov majú v SaS ORG .

„Takto sa nedá pristupovať k verejným peniazom, z večera do rána predložiť nejaký dokument a rýchlo schváliť, a nepýtať sa," povedal minister hospodárstva Richard Sulík PER .

Návrh nepodporil ani jeho stranícky kolega a minister školstva Branislav Gröhling PER .

Sulík povedal, že proti by hlasoval aj minister zahraničných vecí Ivan Korčok PER , ak by bol na rokovaní vlády prítomný.

„Toto je proti DNA SaS ORG , aby sme tu len takto zvýšili výdavky o 3,5 miliardy eur. Je to vec, ktorú chceme prediskutovať, máme tam mnoho otázok ku konkrétnym položkám."

Žiada preto riadnu diskusiu a odôvodnenie.

Nejde však o novú informáciu, že deficit sa má v tomto roku zvýšiť na takmer 10 %. Hovorí sa o tom aj v pláne obnovy, ktorý vláda pred dvoma týždňami poslala do Bruselu LOC .

```
zdroj: www.dennikn.sk
```

Spacy (3.0) + WikiANN

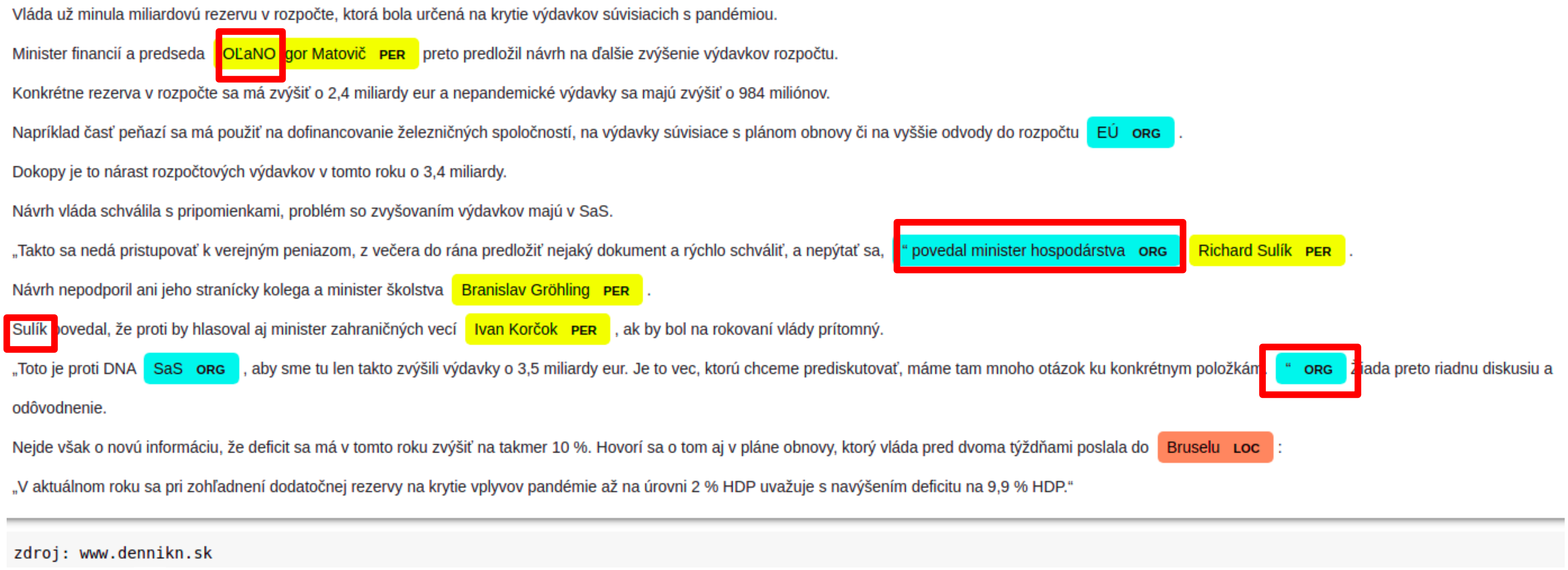
- spacy – NLP library focus on production-ready models
- tok2vec + Transition-based parsing

	Precision	Recall	F1
PER	84.80	88.18	86.46
LOC	86.27	84.46	85.36
ORG	81.55	74.65	77.95
	84.49	82.96	83.72

- embeddings from multilingual BERT + Transition-based parsing

	Precision	Recall	F1
PER	95.71	95.52	95.62
LOC	92.16	93.50	92.82
ORG	89.45	86.91	88.10
	92.64	92.33	92.49

Spacy + WikiANN



Vláda už minula miliardovú rezervu v rozpočte, ktorá bola určená na krytie výdavkov súvisiacich s pandémiou.

Minister financií a predseda OĽaNO Igor Matovič PER preto predložil návrh na ďalšie zvýšenie výdavkov rozpočtu.

Konkrétne rezerva v rozpočte sa má zvýšiť o 2,4 miliardy eur a nepandemické výdavky sa majú zvýšiť o 984 miliónov.

Napríklad časť peňazí sa má použiť na dofinancovanie železničných spoločností, na výdavky súvisiace s plánom obnovy či na vyššie odvody do rozpočtu EÚ ORG .

Dokopy je to nárast rozpočtových výdavkov v tomto roku o 3,4 miliardy.

Návrh vláda schválila s pripomienkami, problém so zvyšovaním výdavkov majú v SaS.

„Takto sa nedá pristupovať k verejným peniazom, z večera do rána predložiť nejaký dokument a rýchlo schváliť, a nepýtať sa, “ povedal minister hospodárstva ORG Richard Sulík PER .

Návrh nepodporil ani jeho stranícky kolega a minister školstva Branislav Gröhling PER .

Sulík povedal, že proti by hlasoval aj minister zahraničných vecí Ivan Korčok PER , ak by bol na rokovaní vlády prítomný.

„Toto je proti DNA SaS ORG , aby sme tu len takto zvýšili výdavky o 3,5 miliardy eur. Je to vec, ktorú chceme prediskutovať, máme tam mnoho otázok ku konkrétnym položkám. “ ORG Žiada preto riadnu diskusiu a odôvodnenie.

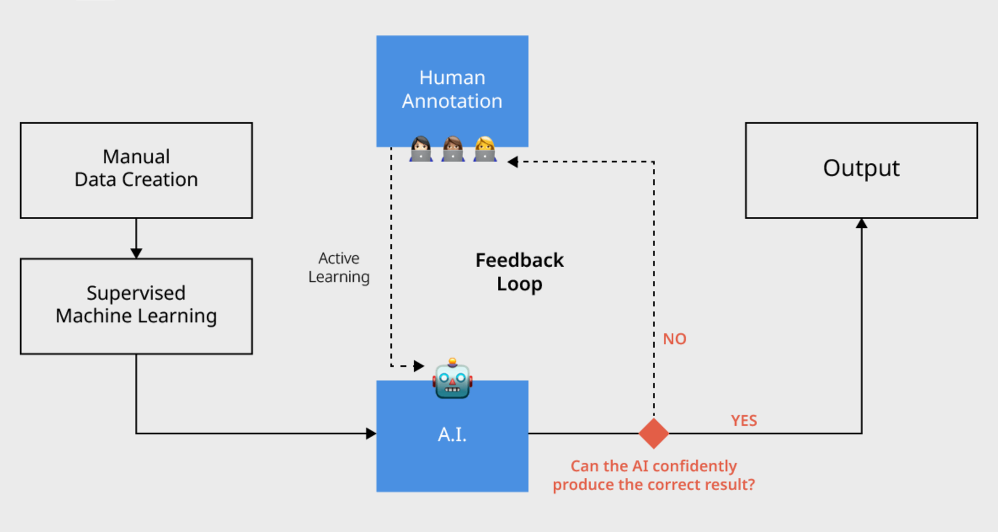
Nejde však o novú informáciu, že deficit sa má v tomto roku zvýšiť na takmer 10 %. Hovorí sa o tom aj v pláne obnovy, ktorý vláda pred dvoma týždňami poslala do Bruselu LOC :

„V aktuálnom roku sa pri zohľadnení dodatočnej rezervy na krytie vplyvov pandémie až na úrovni 2 % HDP uvažuje s navýšením deficitu na 9,9 % HDP.“

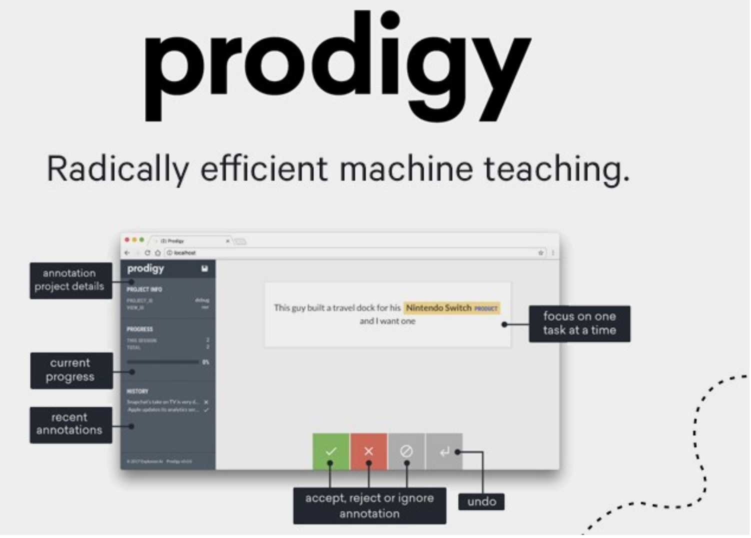
```
zdroj: www.dennikn.sk
```

Human in the loop + Active learning

- annotations which are model uncertain with (or all of them) go to human check
- solution for low resources in slovak NER
- Prodigy (Spacy)



source: https://lionbridge.ai/articles/what-is-human-in-the-loop-machine-learning/

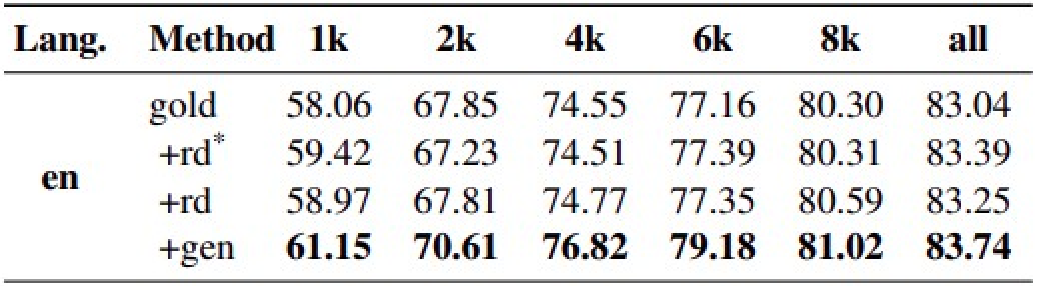

prodigy
Radically efficient machine teaching.
annotation project details
current progress
recent annotations
focus on one task at a time
accept, reject or ignore annotation
undo
PROJECT INFO
PROGRESS
HISTORY

This guy built a travel dock for his Nintendo Switch PRODUCT and I want one

source: https://prodi.gy/

Dataset Augmentation

- regularly used in various taks like computer vision (rotation, cropping…) or speech recognition (noise, pace…)
- method used in NLP e.g. back–translation or random deletion/swap/insertion are much harder to use at token level taged sequntial data
- approach proposed at [4] for tasks with token level tags had interesting results and we would like to try it in our case

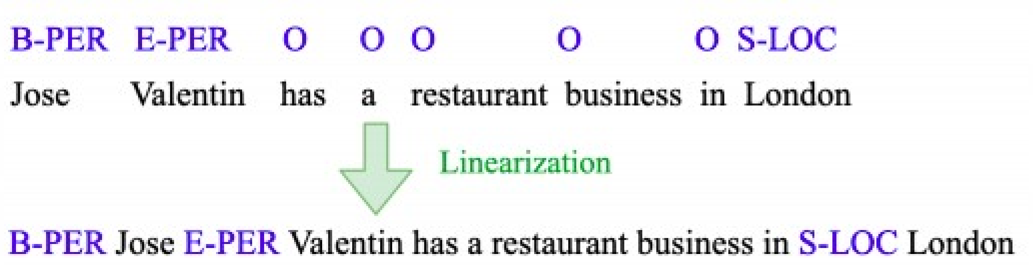


Lang.	**Method**	**1k**	**2k**	**4k**	**6k**	**8k**	**all**
en	gold	58.06	67.85	74.55	77.16	80.30	83.04
	+rd*	59.42	67.23	74.51	77.39	80.31	83.39
	+rd	58.97	67.81	74.77	77.35	80.59	83.25
	+gen	**61.15**	**70.61**	**76.82**	**79.18**	**81.02**	**83.74**

source: [2]

Dataset Augmentation

- Steps:
  - linearization of golden dataset with labels
  - training language model from dataset modified like mentiond
  - generation of syntetic data: language model generates sentence with NER tag (e.g. LOC) after which model can choose from many locations he has seen during training
  - ‘de-linearization’ of generated sentences



B-PER	E-PER	O	O	O	O	O	S-LOC
Jose	Valentin	has	a	restaurant	business	in	London

B-PER Jose E-PER Valentin has a restaurant business in S-LOC London

source: [4]

Literatúra

- [1] Kaššák, Ondrej - Kompan, Michal - Bieliková, Mária. Extrakcia pomenovaných entít pre slovenský jazyk. In ZNALOSTI 2012 : Sborník příspěvků 11 ročníku konference, Mikulov, hotel Eliška 14.-16. 10. 2012. Praha:Matfyzpress, 2012, pp. 52–61

- [2] Ashish Vaswani and Noam Shazeer and Niki Parmar and Jakob Uszkoreit and Llion Jones and Aidan N. Gomez and Lukasz Kaiser and Illia Polosukhin. 2017. Attention Is All You Need. arXiv:1706.03762

- [3] Xiaoman Pan, Boliang Zhang, Jonathan May, Joel Nothman, Kevin Knight, and Heng Ji. 2017. Cross-lingual name tagging and linking for 282 languages. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 1946–1958

- [4] Bosheng Ding, Linlin Liu, Lidong Bing, Canasai Kruengkrai, Thien HaiNguyen, Shafiq Joty, Luo Si, and Chunyan Miao. 2020. DAGA: Data augmentationwith a generation approach for low-resource tagging tasks. InProceedingsof the 2020 Conference on Empirical Methods in Natural Language Process-ing (EMNLP), pages 6045–6057. Association forComputational Linguistics.

Key articles

- David Ifeoluwa Adelani et al. 2021. MasakhaNER: Named Entity Recognition for African Languages. arXiv:2103.11811
- Suppa, Marek and Jariabka, Ondrej. 2021. Benchmarking Pre-trained Language Models for Multilingual NER: TraSpaS at the BSNLP2021 Shared Task. InProceedings of the 8th Workshop on Balto-Slavic Natural Language Process-ing, pages 105–114, Kiyv, Ukraine. Association for Computa-tional Linguistics.
- Afshin Rahimi, Yuan Li, and Trevor Cohn. 2019. Massively multilingual transfer for ner. arXiv:1902.00193

Thank you for your (self) attention!